MULTIMEDIA UNIVERSITY

STUDENT ID NO

MULTIMEDIA UNIVERSITY

FINAL EXAMINATION

TRIMESTER 2, 2015/2016

TMM 3341– DATA MINING AND MACHINE LEARNING SYSTEMS

(All sections / Groups)

3 MARCH 2016
9.00am – 11.00am
(2 Hours)

INSTRUCTIONS TO STUDENT

1. This Question paper consists of 4 pages including cover page.

2. Answer **FIVE** out of **SIX** questions. All questions carry equal marks and the distribution of the marks for each question is given.

3. Please write all your answers in the Answer Booklet provided.

QUESTION 1

a. Describe the steps and their purposes in knowledge discovery from databases (KDD). (4 marks)

b. Why is Data Pre-processing important in data mining task? (2 marks)

c. Suppose that you are employed as a data mining consultant for an Internet search engine company. Describe how data mining can help the company by giving **ONE (1)** example of clustering and **ONE (1)** example of classification techniques. (4 marks)

QUESTION 2

a. Suppose a group of 12 students with the test scores listed as follows: 35, 71, 88, 63, 19, 85, 69, 81, 72, 48, 99, 95. Partition them into four bins by

i. equal-frequency (equi-depth) method
ii. equal-width method
iii. clustering (3 marks)

b. Would simple random sampling (without replacement) be a good approach to sampling? Explain your answer. (3 marks)

c. In any classification technique, we need to separate the data into training and test data. Describe the purpose of separating the data into these 2 sets. (4 marks)

QUESTION 3

a. What is overfitting? (2 marks)

b. Eva and Rashid sampled 2,000 and 200 points from the same population, respectively. They both decide to use a hold-out method by keeping 50% for training and 50% for testing. Eva builds a classifier achieving 90% accuracy; Rashid also builds a classifier achieving 90% accuracy. Both argue about which classifier's performance is closer to the actual error rate. Rashid argues that his classifier is better, while Eva argues that her own is better. Who do you think should win this argument and why? (4 marks)

c. Dr. Zaman has a patient who is very sick. Without further treatment, this patient will die in about 3 months. The only treatment alternative is a risky operation. The patient is expected to live about 1 year if he survives the operation; however, the probability that the patient will not survive the operation is 0.3. Draw a decision tree for this simple decision problem. Show all the probabilities and outcome values. (4 marks)

Continued .....

QUESTION 4

a. Let butter, milk and doughnut be item-sets in a dataset of transactions.

i. Define support {butter} → {milk, doughnut}. (1 mark)

ii. Explain why the support of a rule is a useful information in the analysis of a dataset? (1 mark)

b. Explain **TWO (2)** disadvantages of neural network classification methods in data mining. (2 marks)

c. Consider the following eight points
P1 (2, 2), P2 (1, 14), P3 (10, 7), P4 (1, 11), P5 (3, 4), P6 (11, 8), P7 (4, 3), P8 (12, 9)

The distance function between two points P1 (x1, y1) and P2 (x2, y2) is defined as: $D(P1, P2) = |x1 - x2| + |y1 - y2|$

Take P1, P2 and P7 as initial centroids. Then apply k-means clustering algorithm to calculate the **FIRST** successive position of those centroids. (6 marks)

QUESTION 5

a. There are **FOUR (4)** errors in the following .arff file. Correct the errors. Write only the line numbers and the corrected lines. You do not have to rewrite the entire file. (4 marks)

1	relation weather
2	@attribute outlook {sunny, cloudy, rainy}
3	@attribute temperature numeric
4	@attribute humidity {low, high}
5	@attribute windy {TRUE, FALSE}
6	@attribute play yes, no
7	@data
8	sunny,85,85,FALSE,no
9	sunny,80,90,TRUE,no
10	overcast,83,86,FALSE,yes
11	rainy,70,96,FALSE,yes

b. Suppose a text retrieval system has retrieved a number of documents for a user based on the user's input in the form of query. How can the user assess the quality of the text retrieval system? (4 marks)

c. With reference to your data analysis of personal food tastes, explain how would you use the pattern findings to improve business. (3 marks)

Continued .....

QUESTION 6

a. Define **THREE (3)** roles of Exploratory Data Analysis (EDA) plays in a data mining project? (3 marks)

b. Use the three-class confusion matrix below to answer the questions below.

Computed Decision

	Class 1	**Class 2**	**Class 3**
Class 1	10	5	3
Class 2	5	15	3
Class 3	2	2	5

i. What percent of the instances were correctly classified?
ii. How many *class 2* instances are in the dataset?
iii. How many instances were incorrectly classified with *class 3*?
(3 marks)

c. Given are the following five transactions on items {A;B;C;D;E}

tid	items
1	AB
2	ADE
3	CE
4	BCD
5	ABDE

Use the Apriori algorithm to compute all frequent item sets, and their support, with minimum support 2. Clearly indicate the steps of the algorithm, and the pruning that is performed. (4 marks)

End of Page